Lending Club Case Study

Risk Analytics

Ankita Sethi, 13 Sep 2022

Problem Statement

Save the company from incurring loss which can be due to 2 reasons-

- Giving loans to people who turn out defaulters
- Not giving loans to genuine borrowers

- **In this Case study we will try to analyse the data of the company and find the driver variables which might be the strong indicators of a borrower turning out to be a defaulter**
- **This will help company in making decisions in future whether to lend money or not**

Approach followed

- **Data Understanding**
  - Observing the columns and data fields
  - Understanding the field values through data dictionary
- **Data Cleaning**
  - Dropping Unwanted columns
  - Handling NA values, imputing null values
  - Dropping dubious rows
  - Checking data types of columns
- **Data Analysing**
  - Univariate, bivariate analysis
  - Plotting graphs, deriving relations between variables
- **Results**
  - Results and recommendations based on analysis

```
# importing the required libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

- Imported these libraries for data analysis, numerical computation and plotting graphs

```
# Finding shape(rows and columns) of data set

df_1.shape
```

```
(39717, 111)
```

- Initially there were 111 columns which were difficult to analyse, so we needed to drop few columns which were not relevant in our study

Dropping few columns to simplify analysis- as they are loan behaviour variables, they wont help in analysis¶

21 columns dropped are-

‘delinq_2yrs',
‘earliest_cr_line',
‘inq_last_6mths',
‘open_acc',
‘pub_rec',
‘revol_bal',
‘revol_util',
‘total_acc',
‘out_prncp’,
‘out_prncp_inv’,
'total_pymnt',
‘total_pymnt_inv',
‘total_rec_prncp',
‘total_rec_int',
‘total_rec_late_fee',
‘recoveries',
‘collection_recovery_fee',
‘last_pymnt_d',
'last_pymnt_amnt',
‘last_credit_pull_d',
'application_type'

- **DATA CLEANING**
  - We have deleted rows with value 'Current' in column 'loan_status' as they are ambiguous whether they will default or not
  - Few columns which only have NA as values in all 38577 rows can be dropped. Therefore 49 columns are dropped.
  - Dropped column 'delinq_amnt'and 'acc_now_delinq' as it has a constant value 0 in all rows
  - Dropped column 'policy_code' as it has a constant value 1 in all rows
  - Dropped columns 'grade' and 'sub_grade' as the values in them are assigned by LC
  - Dropped column 'initial_list_status' as it has a constant value 'f' in all rows

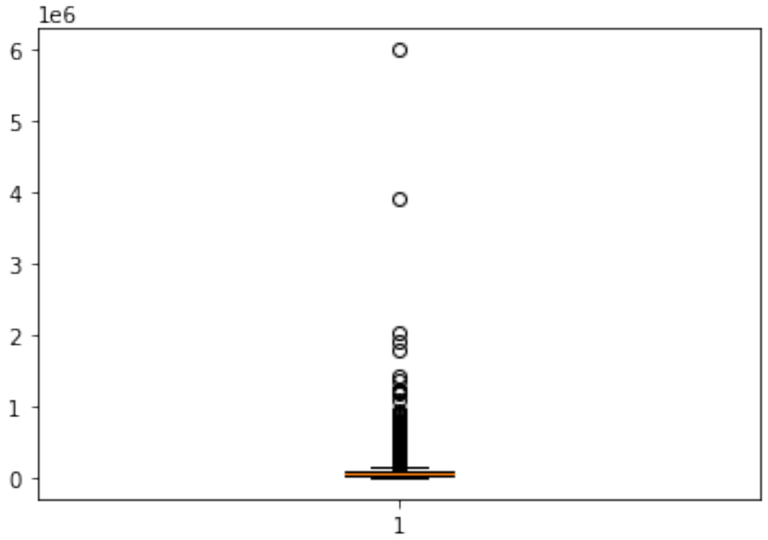
- **SANITY CHECK**
  - funded_amnt_inv' should not be greater than ‘loan_amnt’

- **DATA TYPES CHECK**
  - Make column 'term' int
  - Make column 'int_rate' float

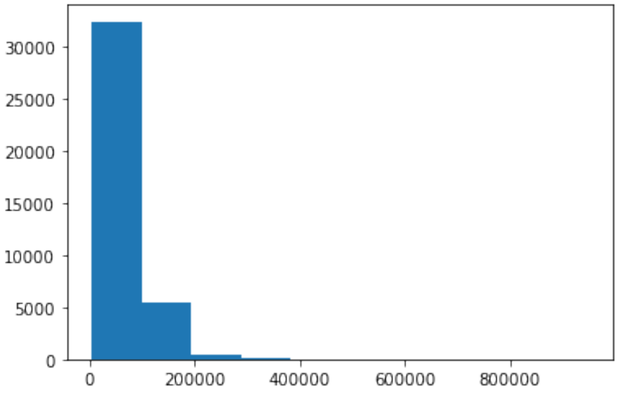
- **DATA ANALYSING**
- **Analysing the column 'annual_inc'**
- FINDINGS:
  - Maximum value seems much higher indicating outliers

```
count    3.857500e+04
mean     6.877584e+04
std      6.421962e+04
min      4.000000e+03
25%      4.000000e+04
50%      5.885256e+04
75%      8.200000e+04
max      6.000000e+06
Name: annual_inc, dtype: float64
```

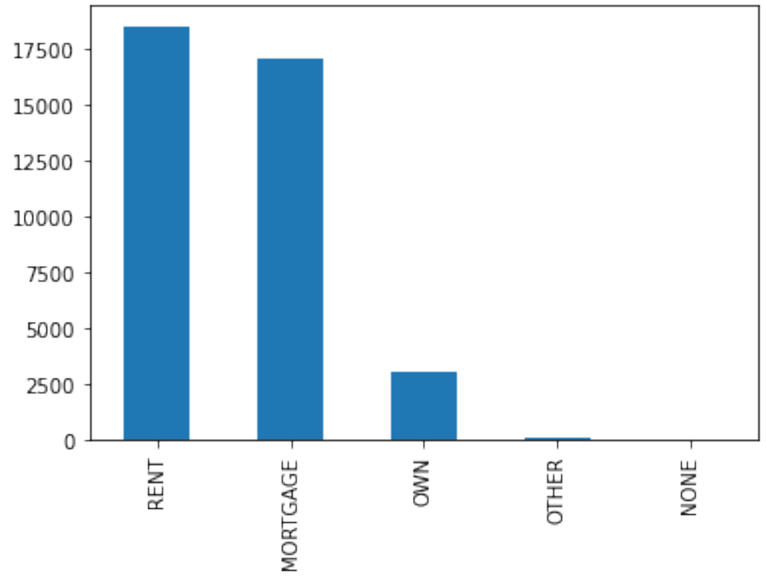
- Boxplot shows there are significant number of outliers. It shows there are signiciant outliers above the upper fence.



- Histogram shows the most of the borrowers have annual income between 0-100000 bucket



- **Analysing the column 'annual_inc'**
- FINDINGS:



- Most of the borrowers taking loan are living on rent or are on mortgage. People who possess their own house borrow less as compared to others